Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web

Ludovic Tanguy & Nabil Hathout
ERSS – CNRS & Université de Toulouse Le Mirail
{tanguy, hathout}@univ-tlse2.fr

Résumé - Abstract

L'article présente Webaffix, un outil d'acquisition de couples de lexèmes morphologiquement apparentés à partir du Web. La méthode utilisé est inductive et indépendante des langues particulières. Webaffix (1) utilise un moteur de recherche pour collecter des formes candidates qui contiennent un suffixe graphémique donné, (2) prédit les bases potentielles de ces candidats et (3) recherche sur le Web des cooccurrences des candidats et de leurs bases prédites. L'outil a été utilisé pour enrichir Verbaction, un lexique de liens entre verbes et noms d'action ou d'événement correspondants. L'article inclut une évaluation des liens morphologiques acquis.

This paper presents Webaffix, a tool for finding pairs of morphologically related words on the Web. The method used is inductive and language-independent. Using the WWW as a corpus, the Webaffix tool detects the occurrences of new derived lexemes based on a given graphemic suffix, proposes a base lexeme, and then performs a compatibility test on the word pairs produced, using the Web again, but as a source of cooccurrences. The resulting pairs of words are used to enrich the Verbaction lexical database, which contains French verbs and their related nominals. The results are described and evaluated.

Mots-clefs – Keywords

Morphologie dérivationnelle, ressource lexicale, Web comme corpus, analogie.
Derivational morphology, lexical resource, Web as corpus, analogy.

1 Introduction

Ce travail s'inscrit dans le cadre général du développement de ressources lexicales pouvant être utilisées dans différentes applications du TAL. L'exemple qui nous a servi de tremplin est le lexique Verbaction[1] qui contient 6 471 couples verbe:nom, tels que le nom est morphologiquement apparenté au verbe et qu'il dénote l'action ou l'événement correspondant à ce verbe. Il s'agit d'une ressource lexicale générique dont des exemples d'utilisation seront présentés ci-dessous. Notre but initial est d'étendre ce lexique, en y ajoutant des liens morphologiques entre

[1]Ce lexique a été réalisé à l'INaLF (CNRS, USR705) à partir de TLFnome, en 1997, par A. Berche, F. Mougin, N. Hathout et J. Lecomte. TLFnome est un lexique de formes fléchies construit à l'INaLF par J. Maucourt et M. Papin, à partir de la nomenclature du *Trésor de la Langue Française* (TLF).

des lexèmes moins courants, voire des néologismes, et en travaillant pour ce faire sur corpus. Mais plutôt que d'exploiter un corpus « traditionnel » comme les archives électroniques de journaux (*Le Monde*, *Libération*, etc.), nous avons préféré utiliser le Web comme corpus. Ce choix, aisément justifiable par la masse d'information textuelle disponible, rend toutefois nécessaire de prendre un certain nombre de précautions, ce corpus étant caractérisé par son hétérogénéité et le manque de contrôle sur son contenu.

2 Vue d'ensemble

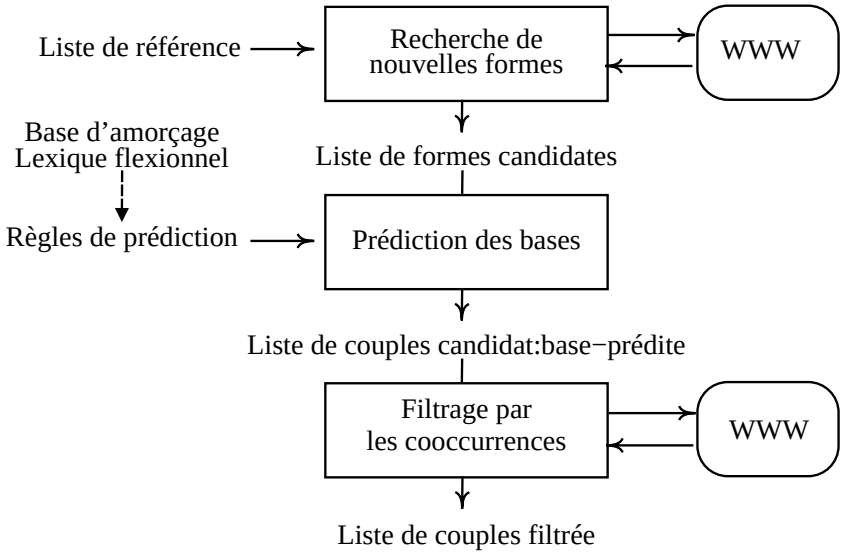

FIG. 1 – Vue d'ensemble des modules de Webaffix.

Notre méthode est représentée graphiquement en figure 1. Elle peut se résumer de la façon suivante : nous recherchons sur le Web, à l'aide d'un moteur de recherche généraliste, des formes lexicales nouvelles, en fonction de leur terminaison (en fait, d'un suffixe graphémique donné). Par « nouvelles », nous entendons simplement absentes d'une liste de référence, par exemple un lexique électronique. Les terminaisons sélectionnées pour l'enrichissement de Verbaction sont les principaux suffixes qui permettent de construire des noms d'action et d'événement à partir de verbes (*-ade*, *-age*, *-ance*, *-ement*, *-ence*, *-erie*, *-tion*). On trouve ainsi par exemple comme formes candidates *dérushage*, *hélitreuillage* pour *-age*, ou *judiciarisation* ou *icônification* pour *-tion*. Une fois ces nouvelles formes collectées, nous calculons pour chacune d'elles des formes graphémiques susceptibles d'être des formes fléchies du lexème base du candidat. Cette prédiction est réalisée par analogie avec une base dérivationnelle existante (comme Verbaction), ou bien à l'aide d'un ensemble de règles d'affixation apprises à partir d'un lexique flexionnel, par exemple, en utilisant la technique proposée dans DéCor (Dal et al. 1999). Dans le cas de notre expérience, les formes candidates sont supposées nominales et les formes bases verbales sont calculées en utilisant Verbaction. Nous prédisons par exemple, pour le candidat *icônification*, les formes *icônifiait*, *icônifiant*, *icônifie*, *icônifient*, *icônifier*, *icônifiera*... de son lexème base. Enfin, dans une troisième étape, nous filtrons les couples lexème-candidat:lexème-base-prédit ainsi constitués en recherchant, toujours sur le Web, à l'aide d'un moteur de recherche, des cooccurrences dans une même page des formes des deux lexèmes du couple candidat. Ne retenant que les couples pour lesquels une telle cooccurrence a été trouvée, nous procédons à une révision manuelle finale, grandement facilitée par le fait que les contextes des formes co-

occurrentes sont conservés par la dernière étape. Notre but n'est pas ici d'obtenir, dans la base de données construite *in fine*, une couverture la plus large possible, mais plutôt de garantir une bonne précision dans les résultats obtenus.

La méthode que nous proposons est assez proche de celle implémentée dans GéDériF (Dal et Namer 2000). Cet outil génère un ensemble de mots possibles candidats à partir d'un lexique puis les filtre pour ne conserver que ceux qui sont attestés en particulier sur le Web. Comme pour Webaffix, le traitement comporte trois étapes : un ensemble de radicaux est constitué en analysant les lemmes du lexique en *-able* et en *-iser* à l'aide de l'analyseur morphologique DériF ; il concatène ensuite à ces lemmes les *-able*, *-iser* et *-ité* (complémentaire avec la terminaison originale du radical) ; GéDériF filtre les candidats obtenus en recherchant leurs attestations éventuelles dans des corpus textuels et sur le Web en utilisant le moteur `www.yahoo.fr`. L'objectif de GéDériF est plus théorique que le nôtre dans la mesure où la construction de lexèmes en *-abiliser*, *-abilité*, *-isable*, *-isabilité*... vise autant à valider les hypothèses linguistiques proposées par les autrices qu'à construire un ensemble de micro-familles constructionnelles.

Il est à noter dès à présent que cette méthode ne demande pour son établissement que peu de ressources : une liste de référence permettant de filtrer les unités déjà répertoriées pour la première étape ; et pour la deuxième étape, un ensemble de règles ou de schémas dérivationnels et flexionnels appris à partir d'une éventuelle base existante et d'un lexique de formes fléchies. Aucune connaissance linguistique n'étant implémentée dans l'outil, il est adapté à l'acquisition d'autres types de liens morphologiques (nom:adjectif, adjectif:adverbe...), mais aussi d'autres langues à morphologie concaténative comme les langues romanes et germaniques. Par ailleurs, cette méthode est incrémentale : Webaffix est destiné à être régulièrement appliqué sur le Web, profitant de l'accumulation de l'information obtenue lors de passages précédents, et de la dynamique des bases de documents indexés par les moteurs de recherche.

3 Le lexique Verbaction

Les ressources que Webaffix permet de constituer sont du type du lexique Verbaction. Cette base est avant tout une ressource pour le TAL destinée au traitement des variations morpho-syntaxiques (Jacquemin 1997). Les hypothèses théoriques sous-jacentes sont minimales. Aucune distinction n'est faite entre les conversions et les constructions suffixales dans la mesure où la nature de converti ou de suffixé n'est pas prise en compte par les systèmes qui utilisent cette ressource (seule compte la relation sémantique entre le verbe et le nom). Les liens sémantiques spécifiques entre noms et verbes ne sont pas explicités. Le principe est ici seulement de répertorier des liens sémantiques entre un verbe et un nom d'action ou d'événement auquel il est dérivationnellement apparenté comme *élire*:*élection*, *déménager*:*déménagement*... Ce type de ressource est utile à divers titres. Nous nous donnons ici quelques exemples de son utilisation.

En recherche d'information, l'équivalence sémantique des variantes morpho-syntaxiques peut aisément être utilisée pour l'extension de requêtes. Par exemple, rechercher *élection présidentielle* est équivalent, en RI, à rechercher *élire le président*.

En analyse syntaxique automatique, certaines relations argumentales peuvent être partagées par les prédicats nominaux et verbaux. L'analyseur syntaxique Syntex (Bourigault et Fabre 2000) utilise dans sa procédure d'apprentissage endogène les liens contenus dans le lexique Verbaction pour repérer ce type de partage. Par exemple, la présence d'un SN comme *des phénomènes de transport sur coussin d'air emprisonné* est un indice en faveur du rattachement de

sur le fond au verbe *transporter* dans *les vagues transportent des sables sur le fond.*[2]

Toutes ces utilisations visent donc à ajouter une annotation supplémentaire d'un corpus (préalablement étiqueté morphosyntaxiquement), et celle-ci sera d'autant plus complète que le lexique recensant ces relations morphologiques sera plus étendu.

4 Le Web comme corpus

Le corpus sur lequel porte notre étude, et notre recherche d'unités et d'informations lexicales, est le Web, et non un corpus classique. Comme le note Grefenstette (1999), nombre de linguistes peuvent, à juste titre, se montrer réticents à utiliser le Web comme source d'attestations, étant donné l'impossibilité technique de caractériser les pages sur le plan du domaine, du genre, du statut de l'auteur, de la validité du contenu, etc. Il n'en reste pas moins que le Web constitue actuellement la masse textuelle accessible la plus importante.

L'utilisation du Web comme corpus se généralise. Signalons, par exemple, le projet WebCorp (`http://www.webcorp.org.uk`) qui met la technologie de base d'un concordancier à l'échelle du Web, en se fiant à plusieurs moteurs de recherche génériques. Ces moteurs restent de toute façon le seul moyen d'accès au Web, à moins de développer un système de parcours et d'indexation spécifique, qui ne pourra de toute façon pas prétendre à l'exhaustivité d'un Google™ et autre AltaVista™. Notre approche n'échappe pas à ce filtre, et nous ne travaillons, comme tous les autres « webolinguistes », que sur la partie du Web (dont la proportion est d'ailleurs inconnue) indexée par ces moteurs, et donc sur un sous-ensemble des pages accessible qui varie avec le temps, et sur lequel aucun critère de sélection fiable n'est applicable.

La question se pose alors du type d'études sur corpus pour lesquelles le recours au Web est justifié. Les études actuelles vont du repérage d'entités nommées (Jacquemin et Bush 2000) à l'étude de textes parallèles bilingues (Resnik 1999), et en règle générale se limitent à l'étude d'unités lexicales en utilisant des extracteurs de contextes et des mesures de cooccurrence. Il paraît en effet plus délicat de mener des études relevant de la syntaxe ou de la sémantique sur un corpus aussi « incontrôlable », et pour lequel on dispose de si peu d'informations.

Nous nous situons ici dans le cadre de la morphologie, avec comme but affiché de constituer des ressources facilement réutilisables. Nous proposons une méthode permettant de construire des bases de données lexicales, dont la couverture va croissante, et dont le contenu peut être supposé indépendant d'un domaine particulier[3]. Puisque nous nous concentrons sur des unités non référencées dans des dictionnaires usuels, les unités que nous recueillons relèvent soit de la création lexicale spontanée et éphémère, soit de langues de spécialité. Dans tous les cas, leur description et leur accumulation, après révision manuelle, constitue une ressource aisément utilisable et utile pour divers traitements des textes en TAL (cf. §3). Notre hypothèse de travail sera donc minimale, et ne suppose qu'un comportement linguistique très général sur le mode de formation et d'usage des unités lexicales construites, indépendante du type de texte et du domaine dans lesquels ces constructions apparaissent (à quelques exceptions près, certains documents disponibles sur le Web ayant un statut particulier, comme nous l'explicitons en §8).

[2] Ces exemples sont empruntés à (Hathout et Fabre 2002).

[3] Il existe un biais dont nous devons prendre conscience : il est illusoire de penser que le Web serait un corpus de « langue générale ». Si la variété des domaines abordés, et donc des sous-langages de spécialité représentés peut paraître suffisante, nous n'avons pas d'idée claire de la représentation de chacun de ces domaines. En attendant la mise en place de procédures génériques de profilage et de caractérisation des pages Web, comme celle proposée que le projet TyPWEB (Beaudouin et al. 2001) nous nous contentons de ce qui est disponible.

5 La cooccurrence des lexèmes construits et des lexèmes bases

(Baayen et Neijt 1997) ont montré que les contextes des mots dérivés contiennent fréquemment des « ancres » c'est-à-dire des indices qui facilitent leur interprétation. C'est ainsi que les mots dérivés apparaissent régulièrement précédés (et plus rarement suivis) d'une forme du lexème sur lequel ils sont construits.

La présence du lexème base dans le contexte peut relever de la coopération conversationnelle : il s'agit alors de fournir des éléments permettant d'interpréter le dérivé en le mettant en relation avec sa base. Si la dérivation morphologique est un moyen d'exprimer des notions complexes de manière concise, on peut néanmoins supposer qu'en français comme dans d'autres langues, elle permet assurer la continuité thématique et référentielle dans le discours mais aussi d'éviter les répétitions et même de varier la façon dont les idées sont présentées et développées. Dans tous ces cas, on peut faire l'hypothèse que les lexèmes construits peuvent être utilisés pour paraphraser leurs lexèmes bases. La cooccurrence des deux lexèmes est ainsi prévisible sans être systématique. Cette observation a été exploitée en RI par (Xu et Croft 1998) pour filtrer des appariements morphologiques produits par un raciniseur. Le filtrage est basé sur une variante de la mesure d'information mutuelle attendue (EMIM ; *Expected Mutual Information Measure*) calculée entre des formes morphologiquement apparentées cooccurrentes dans des fenêtres de 200 mots. En nous appuyant sur cette même observation, nous proposons une technique simple permettant des liens morphologiques, en l'occurrence des couples de lexèmes dérivationnellement apparentés. Il s'agit d'explorer le Web pour y chercher des pages qui contiennent des formes des deux lexèmes qui composent le lien.

Les formes de la base potentielle étant prédites (cf. §7), leur attestation est un premier indice du caractère construit du lexème candidat —Webaffix rejoint sur ce point GéDériF. Par ailleurs, la cooccurrence des formes du lexème candidat et de la base prédite constitue un indice fort de l'existence d'une relation sémantique entre ces deux unités. La nature exacte de cette relation n'est cependant pas connue, même s'il est probable qu'elle corresponde au sens associé à l'affixe dont on a fait l'hypothèse de la présence pour prédire le lexème base.

Dans les trois sections suivantes, nous décrivons en détails la méthode employée par Webaffix pour collecter et filtrer des informations morphologiques dérivationnelle en explorant le Web.

6 Première étape : recherche de nouvelles formes sur le Web

Le but de la première étape est de constituer un ensemble de formes se terminant par un suffixe graphémique donné, et qui n'appartiennent pas un lexique de référence. Pour compléter Verbaction, nous avons utilisé le lexique TLFnome, mais l'on peut aussi partir des listes de formes accessibles sur le Web. Webaffix peut aussi fonctionner sans lexique de référence mais un important travail de dépouillement initial est alors nécessaire ainsi qu'une application répétée de la méthode afin d'obtenir un noyau de base.

Comme pour chaque utilisation du Web comme corpus, il est impératif de prendre en compte les spécificités et les limites des moteurs de recherche. Dans notre cas, la recherche est basée uniquement sur la terminaison des formes. Peu de moteurs proposent cette fonctionnalité, très coûteuse en temps de calcul et d'une utilité jugée faible pour l'usage classique des moteurs par les internautes. À notre connaissance, seuls NorthernLight™ et AltaVista offrent un tel service[4].

[4] `http://www.northernlight.com` et `http://www.altavista.com`

Ces deux moteurs imposent toutefois de spécifier un nombre minimum de lettres à l'initiale de la chaîne recherchée : quatre pour NorthernLight et trois pour Altavista. Par exemple, pour rechercher les formes en *-tion*, on doit soumettre dans le cas d'AltaVista un ensemble de sous-requêtes : `aba*tion`, `abc*tion`, ..., `zyt*tion`. Les trigrammes initiaux peuvent être générés en énumérant les 60 000 combinaisons de trois des caractères du français (lettres accentuées comprises). Une réponse plus économique consiste à se limiter aux seules séquences qui se trouvent à l'initiale des entrées de notre lexique de référence, le nombre de trigrammes est ainsi réduit à 3 500. Les sous-requêtes sont complétées en interdisant les réponses déjà présentes dans notre liste de référence. Par exemple, nous interdisons la présence de la forme *ablation* dans les pages ramenées par la sous-requête `abl*tion`. Ce premier filtrage se fait en utilisant les fonctionnalités du langage de requête du moteur AltaVista[5]. Enfin, nous restreignons la recherche aux pages indexées par le moteur comme étant rédigées en français. Ce dernier critère n'est pas entièrement fiable, comme il sera précisé en §9.1.

Les pages fournies en réponse par le moteur sont ensuite analysées une à une. Pour chaque sous-requête, vingt pages sont considérées au maximum. Elles sont rapatriées et analysées par Webaffix, car le moteur ne fournit pas la forme qui a motivé leur sélection. Un filtrage important doit être effectué sur ces réponses pour éliminer le bruit. Les erreurs ont des origines diverses : fautes de frappe et d'orthographe ; segments de textes en langue étrangère ; noms propres ; adresses de courrier électronique ; segments d'URL ; noms de variables ou de fonctions des langages de scripts ; noms de fichiers ; extraits de code informatique... Des méthodes spécifiques sont employées pour chacun de ces problèmes. Un mini-correcteur orthographique, gérant essentiellement les problèmes d'accentuation, de redoublement de lettres, et de concaténation de mots, a été mis en place. Pour les noms propres et les variables de code, nous nous appuyons sur la casse (nous ne retenons que les mots composés uniquement en minuscule) et sur les caractères environnants la chaîne repérée (`/`, `@`, `$`...). Enfin, pour identifier les pages et les segments rédigés dans des langues autres que le français[6], nous utilisons une procédure de détection basée sur le repérage, dans le voisinage de la forme candidate, de mots-outils de langues « parasites » —actuellement, l'anglais, l'allemand, l'espagnol et l'italien. Par ailleurs, une partie non négligeable des liens retournés par le moteur de recherche sont obsolètes : page inexistante ou ne contenant pas (ou plus) l'unité qui a servi à l'indexer. C'est notamment le cas des périodiques en ligne, dont les pages changent régulièrement de contenu mais pas de référence dans l'index.

Le temps total de calcul nécessaire à la collecte des formes candidates (cf. figure 1) est d'environ 40 heures par suffixe, le facteur limitant étant le débit de la connexion Internet. Toutes les données présentées dans le présent article ont été obtenues en interrogeant AltaVista entre le 28 Janvier et le 4 Février 2002. Le tableau 1 résume les résultats globaux de la première étape. Les opérations de filtrage citées plus haut éliminent environs 60% des formes candidates. La répartition des différentes causes de rejet de ces occurrences[7] est présentée dans le tableau 2.

[5]L'interdiction des formes de TLFnome peut bien sûr entraîner le rejet de certaines occurrences, dans le cas exceptionnel où une forme du lexème construit candidat apparaît dans la même page qu'une forme de la liste de référence ayant les mêmes lettres à l'initiale et en finale. Toutefois, la réduction du rapport signal/bruit est telle que nous préférons courir le risque de manquer ces quelques occurrences pour ne pas avoir à en filtrer par la suite plusieurs dizaines de milliers.

[6]En effet, les moteurs de recherche n'attribuent qu'une seule langue à chaque document indexé, et peuvent ainsi déclarer comme francophones des pages bilingues, ou rédigées en latin, occitan, picard, ancien français...

[7]Dans les cas d'absence du candidat dans la page et de lien obsolète, nous considérons que chaque réponse d'AltaVista concerne une occurrence seulement. Ceci n'est bien sûr pas toujours le cas, certaines pages peuvent contenir plusieurs occurrences de formes répondant au schéma recherché.

Suffixe	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Pages analysées	11 618	22 599	15 132	19 528	9 945	12 951	27 664	120 170
Occurrences retenues	2 507	9 962	4 093	10 857	3 503	2 979	12 162	47 429
Noms candidats	813	2 189	1 097	3 791	999	995	3 564	13 448

TAB. 1 – Résultats globaux du premier module.

Suffixe	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Moyenne
Code informatique	6,0	7,2	12,5	3,2	8,0	2,7	5,2	6,3
Absence de la page	26,7	39,1	37,5	42,7	37,2	35,9	36,9	37,0
Mot en majuscules	57,0	39,8	37,8	34,0	39,5	50,1	40,7	41,9
Langue étrangère	1,8	3,3	3,3	3,5	4,3	0,4	7,0	4,0
Liens obsolètes	8,1	7,9	7,6	10,4	9,5	8,6	8,0	8,4
Faute d'orthographe simple	0,3	2,7	1,4	6,1	1,6	2,4	2,3	2,5

TAB. 2 – Répartition par suffixe des opérations de filtrage (en pourcentage d'occurrences).

7 Prédiction des formes des lexèmes bases

```
(crénelage OR crénelages) AND (crénela OR crénelai OR crénelaient OR
crénelais OR crénelait OR crénelant OR … OR créneler OR crénelez OR
créneliez OR crénelions OR crénelle OR crénellent OR crénellera OR …
OR crénelleront OR crénelles OR crénelons OR … OR crénelé OR crénelée
OR crénelées OR crénelés OR crénèle OR crénèlent OR crénèlera OR … OR
crénèleront OR crénèles)
```

FIG. 2 – Requête construite pour le candidat *crénelage* par analogie avec *agnelage*:*agneler*.

Le lexique Verbaction permet de prédire les formes fléchies des verbes susceptibles d'être les bases des formes candidates fournies par le premier module. La prédiction s'appuie sur un ensemble de schémas de suffixation spécifiques à chaque suffixe, appris en extrayant de Verbaction les couples de lemmes dont le nom se termine par ce suffixe ; on génère ensuite l'ensemble des couples de formes fléchies correspondant en utilisant TLFnome. Les membres de ces couples sont munis de leurs catégories morphosyntaxiques. On applique ensuite une technique similaire à celle de (Dal et al. 1999) : un ensemble de schémas bruts est constitué en supprimant dans chaque couple le préfixe graphémique commun à ses deux formes (le préfixe doit comporter au moins trois caractères) et on retient le schéma brut le plus fréquent. Les schémas retenus sont alors appliqués aux formes candidates. Pour chacune des catégories morphosyntaxique de formes à prédire, et pour chaque forme candidate, on identifie le plus long suffixe de cette dernière qui apparaît dans un schéma de suffixation. On applique ensuite à la forme candidate l'ensemble des schémas qui contiennent ce suffixe. Le choix de travailler directement sur des formes fléchies présente un avantage important : il permet de prédire pour une forme candidate l'ensemble des allomorphies possibles de son éventuel lexème base comme cela est illustré en figure 2.

Une correction des approximations induites par l'apprentissage et l'application des schémas est effectuées lorsque l'on peut identifier le lemme correspondant dans TLFnome. Ainsi, si pour un candidat, trois des formes prédites au moins figurent dans TLFnome comme formes fléchies d'un même lemme, alors on remplace toutes les formes prédites par les formes fléchies de ce lemme. Par ailleurs, on supprime dans tous les cas l'ensemble des formes prédites qui, dans TLFnome, appartiennent à une catégorie non verbale. On garantit ainsi que seules

les formes verbales seront considérées lors du filtrage par les cooccurrences. Cette réduction de l'ensemble des formes prédites a par exemple permis d'éliminer un couple candidat comme *affi-cherie*:*afficher* —*afficherie* est construit sur *affiche*— malgré le fait que les formes du substantif *affiche* soient aussi des formes du verbe *afficher*.

8 Filtrage des couples par recherche de cooccurrences

Suffixe	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Couples recherchés	813	2 189	1 097	3 791	999	995	3 564	13 448
Couples filtrées	55	450	154	385	81	85	611	1 821
Moyenne (%)	6,77	20,56	14,04	10,16	8,11	8,54	17,14	13,54

TAB. 3 – Résultats quantitatifs du filtrage basé sur les cooccurrences.

La troisième étape projette sur le Web les liens construits par le deuxième module pour ne retenir que les couples de formes qui apparaissent simultanément dans au moins une page Web. Cette étape utilise Altavista car ce moteur accepte des requêtes longues, jusque 800 caractères. On ne soumet ainsi qu'une requête par lien, ce qui réduit le temps d'exécution de cette étape[8]. Les pages proposées par le moteur de recherche sont analysées pour vérifier, comme dans la première étape, que les occurrences sont « valides ». Nous présentons dans le tableau 3 les résultats quantitatifs de cette étape pour chaque suffixe étudié.

Notons, lors de ce traitement, l'apparition d'une nouvelle source de bruit, constituée par des pages qui ne sont pas des textes mais des listes de mots : dictionnaires et glossaires en ligne, bases de données lexicales mises à disposition sur le Web par des linguistes informaticiens, listes de mots utilisées respectivement par les pirates informatiques et les responsables de sécurité pour attaquer (resp. protéger) les mots de passe... Ces pages ne sont pas gênantes pour la collecte des formes candidates par le premier module. Elles sont en revanche préjudiciables au bon fonctionnement de notre filtrage par la cooccurrence : la présence simultanée de deux unités lexicales sur des bases purement lexicographiques ne peut pas servir de test linguistique. Nous procédons donc à la détection et au filtrage de ces pages Web, en vérifiant la présence d'unités lexicales graphémiquement proches du candidat, ordonnées suivant l'ordre lexicographique.

9 Évaluation des résultats et discussion

Nous présentons ici une évaluation des résultats de Webaffix, obtenus après un étiquetage manuel des candidats et des couples produits par les des différents modules.

9.1 Collecte des formes candidates

La qualité des résultats à la sortie du premier module a été évaluée manuellement par les auteurs en prélevant un échantillon aléatoire de 100 occurrences pour chaque suffixe. L'évaluation de la qualité du premier module est résumée dans le tableau 4. Sont considérées correctes les formes candidates nominales qui se terminent par le suffixe recherché, même s'il ne s'agit pas d'un déverbal. Si la forme n'est pas un nom commun, elle est étiquetée *mauvaise catégorie* comme

[8]Comme nous l'a signaler un relecteur anonyme, ces requêtes peuvent être scindées en sous-requêtes plus courtes que l'on pourrait soumettre à d'autres moteurs, comme Google.

Suffixe	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Moyenne
Candidat correct (%)	**29**	**66**	**40**	**17**	**34**	**59**	**53**	**45**
Mauvaise catégorie (%)	14	1	5	23	2	10	1	8
Langue étrangère (%)	36	7	14	6	11	11	18	12
Faute d'orthographe (%)	16	23	36	51	50	16	27	33
Code, URL, etc. (%)	5	3	5	3	3	4	1	3

TAB. 4 – Répartition des candidats noms dans les différentes catégories

dans le cas de *vibratoirement* qui est un adverbe[9]. Certains contextes en langues étrangères échappent au test présenté précédemment. Il s'agit de segments trop courts, souvent de citations (comme *Parole d'alcôve et outrageous statement*), de traductions partielles (comme *« Sept jours pour expier » (days of atonement)*), d'ancien français (*oultrageusement*) et divers parlers romans, trop proches du français moderne pour être efficacement filtrés. De même, les fautes d'orthographe résiduelles sont trop complexes pour être traitées efficacement. L'utilisation d'un correcteur orthographique filtre la majorité des créations lexicales que nous recherchons ; cette solution a donc été rejetée. Enfin, les codes informatiques vont de langages de programmation à faible ponctuation à des textes parlant d'informatique (*la fonction comptelement [...]*).

Il est clair qu'une précision moyenne de 45% est assez faible à ce stade, et pourrait être améliorée. Toutefois, le véritable filtrage est effectué par le troisième module, lors de la recherche de cooccurrences. De plus, certaines finales ont plus de propension que d'autres à apparaître dans des contextes erronés : c'est le cas de `ement` qui permet aussi de construire des adverbes, et de `ade` qui apparaît dans de très nombreuses formes espagnoles.

9.2 Évaluation des liens filtrés

Suffixe	*-ade*	*-age*	*-ance*	*-ement*	*-ence*	*-erie*	*-tion*	Total
Couples candidats	55	450	154	385	81	85	611	1 821
Corrects	13	355	31	123	9	14	415	960
Précision (%)	**24**	**79**	**20**	**32**	**11**	**16**	**68**	**52.7**

TAB. 5 – Évaluation du filtrage par les cooccurrences

La qualité des résultats du troisième module a également été évaluée manuellement[10]. Sont considérés comme corrects les liens tels que le lexème candidat est un nom d'action ou d'événement correspondant au verbe prédit. Nous avons aussi accepté les couples dont le candidat n'est pas construit par l'affixe mais l'est par conversion (*cytoponction*:*cytoponctionner*). L'évaluation du filtrage par les cooccurrences est résumée dans le tableau 5.

10 Conclusion

Nous avons présenté une méthode d'acquisition semi-automatique de connaissances dérivationnelles à partir du Web. Cette méthode comporte trois phases : la collecte sur le Web de formes

[9] L'utilisation d'un étiqueteur morphosyntaxique serait trop lourde pour la masse de données à traiter, et les étiqueteurs étant de toute façon peu armés pour catégoriser efficacement les néologismes que nous recherchons.

[10] Pour les suffixes *-age*, *-ance*, *-ement* et *-tion* les valeurs présentées sont extrapolées à partir de l'évaluation d'échantillons aléatoires de 100 lexèmes candidats.

candidates qui comportent un suffixe graphémique ; la prédiction des formes des bases que ces candidats pourraient avoir ; le filtrage par la cooccurrence des couples ainsi constitués. L'article présente une évaluation des résultats centrée sur la précision : elle constitue notre objectif premier.

Notre méthode présente plusieurs points forts que nous envisageons de développer prochainement : elle ne met en œuvre aucune connaissance linguistique, ce qui permet de l'utiliser directement pour d'autres schémas dérivationnels comme adjectif:adverbe, nom:adjectif... et d'autres langues, par exemple pour compléter la partie anglaise de la base morphologique CELEX (Baayen et al. 1995). Cette méthode pourrait également servir de base à une approche plus intuitive et moins applicative pour caractériser les pages Web en fonction des suffixes qu'elles contiennent (en s'intéressant en particulier à l'effet de sensibilisation aux règles ou *rule priming*), en vue de dégager des affinités entre certains paramètres extra-linguistiques et la création lexicale.

Remerciements

Nous remercions Marc Plénat (ERSS, CNRS & Université Toulouse le Mirail) pour l'idée qui est à l'origine de Webaffix, ainsi que Natalia Grabar (DIAM, SIM/AP-HP & Université Paris 6) pour ses conseils techniques.

Références

Baayen, R. H., Neijt, A. (1997), Productivity in context: a case study of a Dutch suffix, *Linguistics*, Vol. 35, 565-587.

Baayen, R. H., Piepenbrock, R., Gulikers, L. (1995), *The CELEX Lexical Database (Release 2)*, CD-ROM, Linguistic Data Consortium, University of Pennsylvania, USA.

Beaudouin, V., Fleury, S., Habert, B., Illiouz, G., Licoppe, C., Pasquier, M. (2001), TyPWeb : Décrire la Toile pour mieux comprendre les parcours. Actes de *CIUST'01 : Colloque International sur les Usages et les Services de Télécommunications*, Paris.

Bourigault, D., Fabre, C. (2000), Approche linguistique pour l'analyse linguistique de corpus, *Cahiers de Grammaire*, Vol. 25, 131-151.

Dal, G., Hathout, N., Namer, F. (1999), Construire un lexique dérivationnel : théorie et réalisation, Actes de la *6e conférence sur le Traitement Automatique des Langues Naturelle*, 115-124.

Dal, G., Namer, F. (2000), Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations, *T.A.L.*, Vol. 41, Num. 2, 423-446.

Grefenstette, G. (1999), The WWW as a Resource for Example-Based MT Tasks, Actes de *ASLIB 'Translating and the Computer' Conference*, London.

Hathout, N., Fabre, C. (2002), *Constitution et exploitation de lexiques de formes déverbales*, Journées d'études sur les déverbaux, SILEX, Lille.

Jacquemin, C. (1997), *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches, Université de Nantes.

Jacquemin, C., Bush, C. (2000), Fouille du Web pour la collecte d'entités nommées, Actes de la *7e Conférence sur le Traitement Automatique des Langues Naturelles*.

Resnik, P. (1999), Mining the Web for bilingual text, Acte de *37th Meeting of ACL*, 527-534, Maryland, USA.

Xu, J., Croft, W. B. (1998). Corpus-Based Stemming using Co-occurrence of Word Variants, *ACM Transaction on Information Systems*, Vol. 16, Num. 1, 61-81.